ПРИНЦИПЫ СОЗДАНИЯ ПОИСКОВОГО ОБРАЗА АНГЛОЯЗЫЧНЫХ ПУБЛИЦИСТИЧЕСКИХ ТЕКСТОВ

Журавкова Е.С.

Белорусский государственный университет, г. Минск

Неотъемлемыми компонентами текста являются его смысл и содержание, которые неразрывно связаны между собой. Несмотря на то, что понятие текста давно знакомое и привычное, в лингвистике текста среди ученых существуют разные точки зрения при определении смысла и содержания текста.

В данной работе мы предприняли попытку рассмотреть принципы создания поискового образа англоязычных публицистических текстов.

Сам процесс описания смыслового содержания текста средствами информационно-поискового языка (ИПЯ) называется *индексирование*м. А именно это процесс присвоения документам и запросам или их частям индексов. Итоговая продукция этих процессов содержится в поисковом образе документа (ПОД) [2].

Поисковый образ документа - поисковый образ, выражающий основное смысловое содержание документа. Поисковый образ документа содержит [5]:

- признаки, необходимые для поиска документа по запросу;
- идентифицирующие и другие сведения о документе: выходные данные, тип, язык и т.д.

Процесс индексирования включает всего несколько главных и второстепенных операций, которые выполняются последовательно или одновременно. Однако число таких операций у разных исследователей и в разных системах может существенно различаться.

Однако основными операциями все же являются [2]:

1. Анализ содержания документа и выбор из текста номинативных лексических единиц, существенных с точки зрения его содержания.

2. Формирование перечня ключевых слов (морфологически нормализованных лексических единиц), используемых в процессе свободного координатного индексирования.

3. Избыточное индексирование - введение в дополнительных лексических единиц, связанных по смыслу с исходными ключевыми словами.

Как работают системы идексирования можно ознакомиться, воспользовавшись локальным вариантом системы Altavista, или обратиться к режиму индексирования реальных информационных систем [5].

Индексирование - одно из основных понятий информационного поиска. Цель данного процесса в информационной системе - приписать каждому документу и запросу некоторое множество «индексов» - идентификаторов, отражающих содержание документа. Идентификаторы называют индексами, рубриками, индексационными терминами, ключевыми словами, дескрипторами, полями и т. д., в зависимости от типа используемого ИПЯ. Эти «индексы» отражают содержание документа и управляют поиском, приводя к тем документам, термины которых оказываются наиболее сходными с терминами запроса [2].

Для определения основного содержания каждого из текстов мы использовали статистический метод. В основе этого метода лежит критерий выделения ключевых или опорных слов, которые пронизывают текст и влияют на понимание всего текста.

Процесс выделения опорных слов каждого анализируемого текста включает в себя следующие этапы:

1) построение частотно-алфавитного словаря для каждого анализируемого текста. В частотно-алфавитном словаре для каждого слова текста указывается, сколько раз оно встретилось в данном тексте, в каком количестве абзацев, в каких абзацах анализируемого текста, а также сколько раз в каждом из данных абзацев. В данной работе мы использовали программу «DICT», чтобы получить частотно-алфавитные словари по анализируемому материалу. Для начала, мы привели отобранные тексты к виду,

соответствующему требованиям, предъявляемым данной компьютерной программой: обозначили начало каждого абзаца символом «*», элимировали переносы, преобразовали строчные символы в прописные символы;

2) выделение из частотно-алфавитного словаря потенциальных опорных слов. Далее они служат основой для деления опорных слов на главные и второстепенные, которые и определяют, в конечном счёте, основное содержание анализируемого текста. Для осуществления данной операции мы использовали компьютерную программу «UNIFY», которая выделяет из частотно-алфавитного словаря слова и словосочетания, имеющие частоту два и более.

Данная операция заключалась в выполнении следующей последовательности действий:

a) удаление из полученных частотно-алфавитных словарей всех служебных слов (предлогов, артиклей, местоимений и так далее);

b) объединение всех грамматических форм одного слова в начальную, словарную форму, то есть все личные формы глаголов сводились к инфинитиву, а существительные – к форме единственного числа именительного падежа;

c) пополнение базы данных потенциальных опорных слов словами схожими по тематике;

d) указание в базе данных возможных словоформ опорных слов.

С учетом того, что английский язык – аналитический язык, является целесообразным построить словарь словоформ для всех входящих в базу данных опорных слов.

Затем полученные списки потенциальных опорных слов были дополнены словами схожей тематики на основании интуитивного и контекстологического критериев. Последним этапом составления словаря опорных слов по тематикам стало указание словоформ для слов, входящих в базу данных.

Особенностью проведенного исследования является создание универсальной системы индексирования англоязычных публицистических текстов аналитических и информационных жанров.